Journal of Scientific & Industrial Research
Vol. 81, June 2022, pp. 643-650

A Novel Method for Lip Movement Detection using Deep Neural Network

Kanagala Srilakshmi[1] and Karthik R[2]*

[1]School of Electronics Engineering, [2]Centre for Cyber Physical Systems and School of Electronics Engineering,
Vellore Institute of Technology, Chennai 600 127, Tamil Nadu, India

Received 16 August 2021; revised 13 May 2022; accepted 13 May 2022

Recognition of Lip movements has become one of the most challenging tasks and has crucial applications in the contemporary scenario. It is the recognition of the speech uttered by individual using visual cues. Visual interpretation of lip movement is especially useful in scenarios like video surveillance, where auditory signals are either not available or too noisy for interpretation. It is also useful for hearing-impaired individuals where audio signal is of no use. Many developments have taken place in this nascent field using various deep learning-based techniques. This research does analysis over various state-of-the-art deep-learning models on MIRACL-VC1 dataset. This study also aims to find out the optimal baseline architecture suitable for building a new model with high accuracy for lip movement detection. The models are trained from scratch over the pre-processed MIRACL-VC1 dataset consisting of small-size images. Experimental observations with state-of-the-art deep learning models indicate that EfficientNet B0 architecture yielded an accuracy of 80.13%. Thus, EfficientNet B0 is further utilized as baseline deep architecture to design a customized model for effective detection. This research proposes an attention based deep learning model combined with Long Short-Term Memory (LSTM) layer having EfficientNet B0 as the backbone architecture. The proposed model yielded an accuracy of 91.13%.

Keywords: Attention, CNN, EfficientNet, Lip movement, LSTM, MIRACL-VC1

Introduction

Interpretation of speech enunciated by a speaker through lip movement is a challenging task that has many applications. In video surveillance, the audio may not be available at all situations, or it is too noisy which makes it difficult to comprehend what exactly the conveyed speech is about. It is also very essential for people with hearing impairment to achieve accurate interpretation of speech. People with hearing impairment are at a higher risk of mental problems than normal people. Hence, efficient listening is crucial to function efficiently in society. Moreover, lip movement detection helps in understanding and reconstruction of speech in noisy environments. Hence, grasping lip movements through visual landmarks without the help of speech is essential in such critical circumstances.

Reading of lip movement is a difficult task due to the varying speeds of talking of the same word or sentence for every individual. Lip movement of the same word also changes with the change in the accent of a speaker and facial features. Many works in this field have been proposed in recent years that principally utilize neural connections to segregate the utterances.

Hence, development of an effective system for lip movement detection is very much essential. In this study, analysis is done over state-of-the-art models such as AlexNet, VGG16, InceptionV1, ResNet50, Auto-Encoder and EfficientNet B0 and comparison with Q-Learning Adaptive Deep Belief Network (Q-ADBN) method that was proposed by Qiao *et al.* over MIRACL-VC1 dataset.[1] It is a diverse and balanced dataset consisting of sequences of lip images. Training is done from scratch over small-sized image sequences. This helps to give efficient discernment towards how well the model is performing for the given dataset. A new model is proposed as a part of this research where Attention and Long Short-Term Memory (LSTM) components are coupled with EfficientNet B0 as base architecture to achieve a better performing and reliable model.

Related Works

This section briefly enlists the existing work done on lip movement recognition using various approaches. Initially, methods such as Hidden Markov Model (HMM), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Linear Discriminant

*Author for Correspondence
E-mail: r.karthik@vit.ac.in

Analysis (LDA), Gaussian Mixture Modeling (GMM), and Random Forest are used to extract, learn, and segregate visual features. HMM was used by Rekik *et al.* to perform lip reading by utilizing the depth of data in a picture.[2] This model pulls out a 3D mouth locale from the 3D face model followed by catching the movement and appearance-based descriptor. In another research conducted by Rekik *et al*., KNN and SVM are used for classification using the same characteristics.[3] An automated lip-reading model is developed by Gergen *et al.* that applies LDA in a HMM and GMM framework over Ground Re-IDentification (GRID) dataset.[4,5] Random Forest Manifold model is considered for training the features by Pie *et al.*[6]

In recent times, many deep learning-based approaches have been used for lip reading with an increase in access to (Graphic Processing Unit) GPU, thus producing cutting edge results. Ngiam *et al.* used Restricted Boltzmann machines independently for sound and video.[7] This model employs deep learning concepts for various modes of data namely audio and video speech identification. The famous deep learning model for lip reading, LipNet, which is proposed by Assael *et al.* makes an initial attempt to do end-to-end sentence-level modeling that learns the sequence model and Spatio-temporal visual elements together.[8] It uses a recurrent neural network framework to understand the text from varying lengths of a sequence of images. Gutierrez *et al.* explore lip reading using a basic combination of CNN and LSTM to train lip image sequences.[9] It also experiments with an ImageNet pre-trained based VGG16 model along with an LSTM layer. Watch, Listen, Attend, Spell (WLAS) method is proposed by Chung *et al.* to understand the movement of the lips and interpret the speech enunciated by the speaker.[10] Additionally, a training scheme is presented to speed up the process of learning. Dataset is constructed from British Television for lip movement visual speech recognition to demonstrate the importance of visual information for speech detection. Afouras *et al.* utilized a feature extractor namely 3D-CNN and ResNet combined with classifiers like Bidirectional-LSTM (Bi-LSTM) and Language Model.[11] Wand *et al.* proposed a model that combines Feed-forward network as a feature extractor and the LSTM layer as a classifier.[12] Shillingford *et al.* utilized 3D-CNN for features extraction and Bi-LSTM along with a Finite-state transducer for the classification of sentences.[13] Wang proposed 3D-CNN with deep layered Bidirectional Convolutional LSTM (Bi-Conv-LSTM) for the classification of words.[14] Petridis *et al.* used 3D-CNN and ResNet and combined with Bi-GRU as a classifier.[15] Other feature extractors such as Depthwise CNN and Attention Encoder combined with Language Model is also used for this study.

Significance of the Proposed Method

Given that speech comprehension using only visual elements like lip movement have many applications, it is imperative to obtain accurate results. Hence, it is very important to analyze existing deep learning architectures and other methods proposed by other researchers to device a method that could achieve optimal performance.

This research not only provides a deep insight into the performance of state-of-the-art deep-learning based models like AlexNet, VGG16, InceptionV1, ResNet50, Auto-Encoder and EfficientNet B0 but also compared with other well performing models like Q-ADBN to explore beyond the existing architectures and attain better understanding of which model achieves best performance. This type of rigorous analysis assures that the proposed model is very robust for innumerable scenarios where lip movement detection is extremely crucial.

Based on the analysis with different state-of-the-art architectures, EfficientNet B0 gave the best performance. This architecture is further refined by integrating it with attention mechanism and LSTM layer.

Materials and Methods

Data Collection and Preprocessing

MIRACL-VC1 dataset is used in the proposed research work.[16] This dataset comprises ten words and ten phrases uttered by ten female speakers and five male speakers. Each word/phrase is spoken ten times. This dataset is rearranged into 20 folders with each folder representing the word/phrase is given in Table 1. Each folder consists of the respective word/phrase uttered by all the speakers pooled together in a single folder. This helps to create a more diverse dataset consisting of male and female speakers producing words at diverse speeds where some speakers have a higher degree of lip movement of the same word compared to others. The dataset has 3000 instances of word and phrases where each folder consists of 150 instances. The rearranged dataset consists of the faces of the speakers speaking the words/phrases.

Table 1 — Corresponding ID and Word/Phrase in the rearranged dataset

Folder ID	Word/Phrase	Folder ID	Word/Phrase
1	Begin	11	Stop navigation
2	Choose	12	Excuse me
3	Connection	13	I am sorry
4	Navigation	14	Thank you
5	Next	15	Good bye
6	Previous	16	I love this game
7	Start	17	Nice to meet you
8	Stop	18	You are welcome
9	Hello	19	How are you?
10	Web	20	Have a good time

It is ideal to have lip images of the speaker to make the detection and classification more efficient. Dlib facial detector is used to detect the face of the speaker and Dlib mouth detector to extract the lips from the image.[17] Since the number of lip images varies for every word/phrase instance data padding is essential. Number 30 is fixed as the maximum length of a word/phrase instance. If the number of lip images is less than 30 then data padding is done by adding blank images to make the number of images as 30 for all word/phrase instances. All the images in the dataset are resized to 32 × 32 × 3 size to maintain uniformity. The dataset is shuffled randomly and converted to Numpy arrays and stored to make computing operations over different deep learning models swiftly. The dataset is divided to training and testing dataset. The training dataset is used to perform with models like AlexNet, VGG16, InceptionV1, ResNet50, Auto-Encoder, Q-ADBN, and EfficientNetB0 network. Then the trained model is evaluated over the testing dataset using various evaluation metrics. Each model's evaluation metrics are collected and aggregated for comparison and discussion. The overall flow of proposed system is depicted in Fig. 1.

Proposed System

The architecture of the proposed model is depicted in Fig. 2. EfficientNetB0 is the backbone of the architecture where Attention block and LSTM layer are added to achieve a finer performance. The architecture of the proposed model, EfficientNetB0 model and Attention block are explained in great depth in this section.

The input for the proposed model in Fig. 2 is the preprocessed MIRACL-VC1 dataset where series of lip images to be categorized into one of the 20 categories. This input is passed on to the time

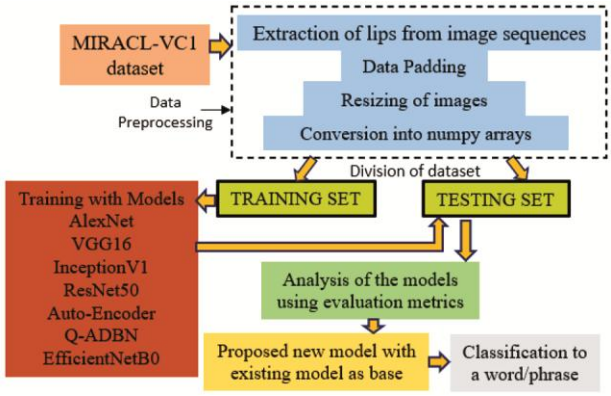

Fig. 1 — Proposed methodology

distributed layer that encloses of EfficientNet B0 model combined with Attention block. Time-Distributed layer is useful to work with time series data, since the data considered for the research consists of series of lip images. LSTM layer accompanied by a Dense Layer is given after the Time-Distributed Layer. LSTM is like that of GRU which was developed from Recurrent Neural Networks (RNN) to combat the vanishing gradient problem, but they differ in architecture. GRU performs better than LSTM in terms of memory and speed. But LSTM gives more accurate results when dealing with datasets having longer sequences. Hence, using LSTM is more beneficial as the pre processed dataset consists of large amount of long sequences of lip images. The final classification of the input series of images is done by Softmax classifier.

Efficient Net Model

EfficientNet model is initially proposed by Tan *et al.* to study and observe the model scaling.[18] It identified that striking a right balance between width, depth and resolution in a neural network can improve the performance drastically. Hence, this study proposed a new method that scales depth, width, and resolution homogeneously. As a part of their study, a new baseline network is used and scaled up to achieve a family of deep learning models such as EfficientNet B0, EfficientNet B1, EfficientNet B2, EfficientNet B3, EfficientNet B4, EfficientNet B5, EfficientNet B6 and EfficientNet B7. These families of EfficientNets attained a superior accuracy as opposed to the existing Convolutional Neural Networks. For this research, EfficientNet B0 is chosen from the existing group of EfficientNets. The architecture of EfficientNet B0 is detailed in Fig. 3. EfficientNet B0 takes in an image of size 32 × 32 × 3 as input. The model is comprised

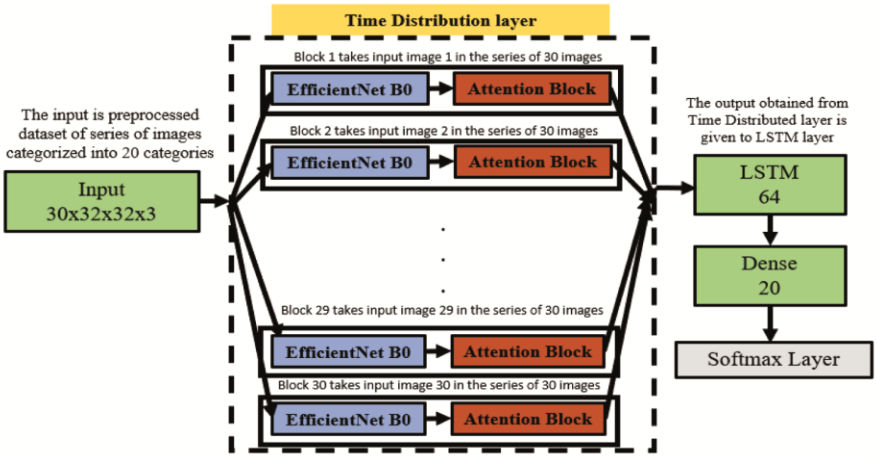

Fig. 2 — Architecture of the proposed model

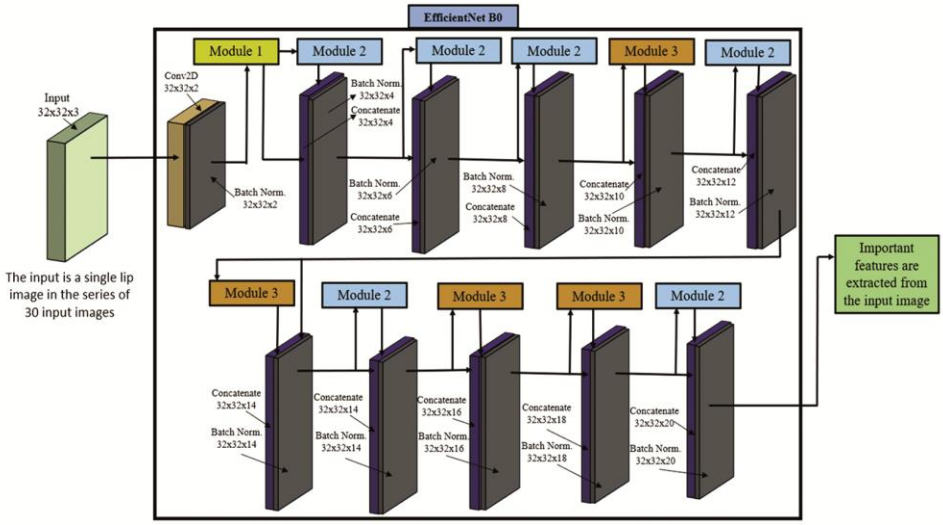

Fig. 3 — Architecture of the EfficientNet B0 model

of 3 modules as shown in Fig. 4 as the main building blocks. It consists of one component of Module 1, six Module 2 components and four Module 3 components.

Module 1 consists of a Depthwise Convolutional 2D layer accompanied by a Batch Normalization layer. Depthwise convolution has a single convolutional filter for every input channel. Unlike regular 2D convolution, in Depthwise convolution each input is convolved with its respective filter and stacked up the convolved outputs together at the end. Batch normalization layer is used to normalize the inputs given to a network and hence useful to give good results. Module 2 consists of four Depthwise Convolutional 2D layers accompanied by a Batch Normalization layer. After two Depthwise Convolutional 2D layers there is a 2D Global Average Pooling coupled with a Rescaling layer. Global Average pooling layer generates one feature map, an average of every feature map is considered as input for next layer. Global Average pooling layer overcomes the overfitting issue since no parameters are required. It also sums up the spatial information which in turn gives stable translations for the given input. Module 3 consists of a Global Average pooling 2D layer combined with Rescaling layer. After rescaling layer there are 2 Depthwise Convolutional 2D layers with every Depthwise Convolutional layer accompanied by a Batch Normalization layer.

Attention Block

Attention block is used to given attention to a particular part while processing the input. The architecture of the attention block used is depicted in great depth in Fig 5. The Attention block takes in

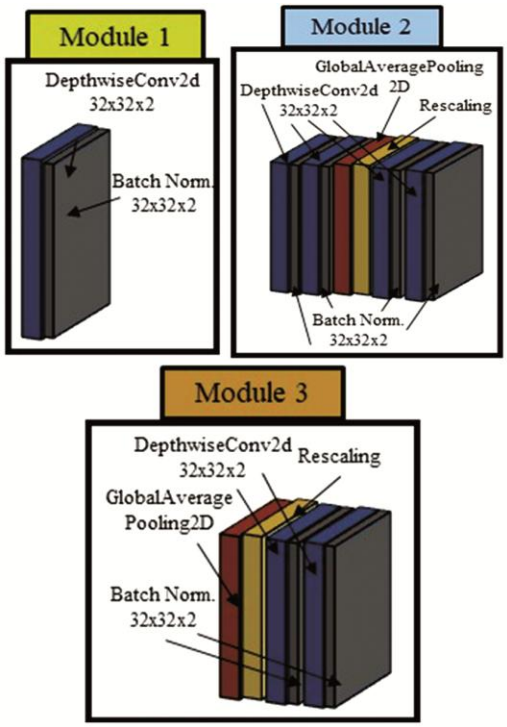

Fig. 4 — Architecture of the Modules used in the EfficientNet B0

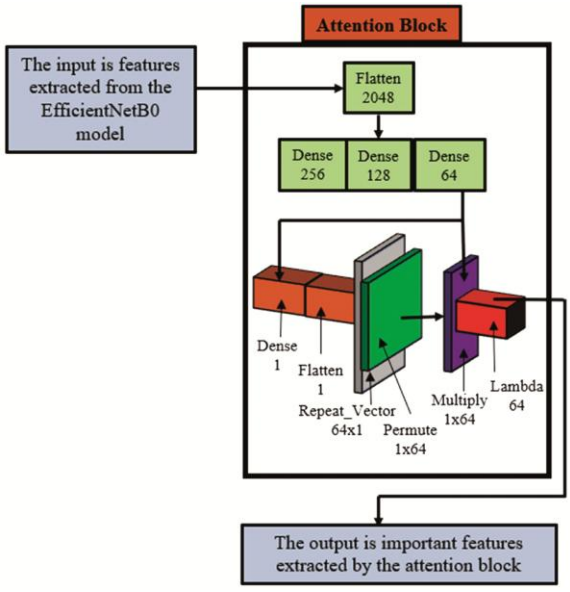

Fig. 5 — Architecture of the Attention Block

input as the features extracted by the EfficientNet B0 neural model.

The main components of Attention Block are Repute Vector, Permute and Lambda. Repute vector is used to add an additional dimension to the given input. Permute layer is used to alter the input shape of the input given to this layer. The Lambda layer is best suitable for simple operations and swift experimentation. The output from the Lambda layer is taken as the output from the attention block.

Results and Discussion

An in-depth analysis is performed over various deep learning models using performance metrics such as Accuracy, Precision, Recall and F1-score as a part of this study. The results obtained have been useful to propose a new deep learning classification method for achieving a better performance. The performance analysis of all models used in this research is explained in this section.

Environmental Setup

The deep learning models used and proposed for this research are trained over Tesla K80 GPU which has a Base Clock Speed of 560 MHz, Boost Clock Speed of 875 MHz, and CUDA Cores as Stream Processors. During the training of a model, a checkpoint is added with validation accuracy as monitor, patience as 20, save_best_only parameter as True, and the number of epochs as 200. When there is no improvement in validation accuracy of the model compared to the last 20 epochs then the training automatically comes to a halt. The weights learned during this process are saved in a .h5 format file to fit it and check over the test dataset and judge the performance of the model. The Optimizer used, Learning rate, and Loss used for all the State-of-the-Art models are Adam, 0.001, and categorical cross-entropy respectively.

Ablation Studies

Analysis of Backbone Architecture

An analysis is done over various State-of-the-Art deep learning-based models like AlexNet, VGG16, InceptionV1, ResNet50, Auto-Encoder and EfficientNetB0. AlexNet is initially proposed by Krizhevsky *et al.*[19] to train large number of high-resolution images of size 224 × 224 over the ImageNet dataset. Simonyan *et al.* proposed the famous deeply connected neural network VGG16 and submitted it for the ILSVRC competition held in 2014.[(20)] It is an improvement over AlexNet since it replaces large kernel size filters with 3 × 3 size kernel filters. InceptionV1 neural network is initially pitched in by Szegedy *et al.*[21] It put forth a breakthrough performance in the ILSVRC competition in 2014 and was declared the winner. InceptionV1 is a sparsely connected architecture that combats the overfitting issues in a deeply connected convolutional

architecture and high requirement of computational resources with an increase in the number of parameters. ResNet50 is proposed by He *et al.*[22] ResNet50 is a 50 layered deep neural network that uses the concept of Skip Connection. Auto-Encoder is proposed in the paper titled "Reducing the Dimensionality of Data with Neural Networks" by *Hinton et al.*[23] The main aim of Auto-Encoder is to try to produce an image that is as close to the given input image. This process can help in dimensionality reduction and removal of noise in the input image. Qiao *et al.* have created the Q-ADBN algorithm for the handwritten digit recognition.[1] This algorithm is created to improve the recognition accuracy and running time by adapting deep Q-learning strategy with the reinforcement learning to form an adaptive Q-learning deep belief network (Q-ADBN). The results in this paper indicate that the proposed Q-ADBN has a superiority to other similar methods in terms of accuracy and running time.

From Table 2, it can be inferred that only AlexNet and Q-ADBN yielded accuracy around 70%. VGG16 yielded accuracy of about 58%. Other models such as InceptionV1, ResNet50 and Auto-Encoder have yielded accuracy of below 50%. EfficientNet B0 has put forth the best performance compared to the other deep learning architectures used for the purpose of the research, giving an accuracy of about 80%. EfficientNet B0 neural network has good equilibrium between evaluation metrics such as Precision, Recall and F1-Score where all its values are close to 1. Therefore, EfficientNet B0 is used as the base architecture for improving the detection of lip movement.

Table 2 — Analysis of performance for State-of-the-Art models

Model Name	Accuracy in %	Precision in %	Recall in %	F1-Score in %
AlexNet	69.12	71.27	69.90	68.34
VGG16	57.23	58.67	57.87	56.334
InceptionV1	28.56	21.789	28.894	21.897
ResNet50	43.89	27.89	43.56	32.67
Auto-Encoder	47.56	48.344	47.234	47.76
Q-ADBN	73.781	72.998	73.87	72.675
EfficientNetB0	80.133	81.89	80.67	80.34

Analysis of Proposed Experiments

From Table 3, it can be observed that a study is performed using Attention Block and LSTM with EfficientNet B0 as backbone architecture. An accuracy of 88.67% is obtained after training over the preprocessed dataset when EfficientNet B0 is combined with Attention Block. The accuracy achieved is around 6% higher than the accuracy achieved by using EfficientNet B0 alone. When LSTM layer is added on top of EfficientNet B0 and Attention Block, the accuracy crossed over 90 percent. Hence, the performance of this model was boosted by 4.46%. Overall, there was an increase in accuracy by 10.997% with addition of Attention block and LSTM over EfficientNet B0 neural network. Hence, this model has yielded more robust performance when compared to the other state-of-the-art deep learning architectures.

Cross Validation Performance for the Proposed Model

The proposed deep neural network has base architecture as EfficientNet B0 and Attention block with LSTM layer additionally added to it. This neural network is trained over the preprocessed dataset. The dataset is divided into training, testing, and validation dataset in 80:10:10 ratio respectively. Five-fold cross-validation is performed where each fold ran for 93 epochs as explained in Fig. 6. The Accuracy,

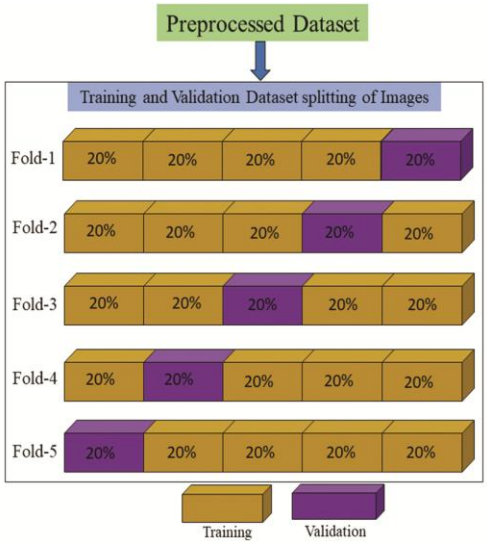

Fig. 6 — Cross-Validation performed for 5 folds over the pre-processed dataset

Table 3 — Analysis of performance for EfficientNet B0, EfficientNet B0 with attention and EfficientNet B0 with attention and LSTM

Model Name	Accuracy in %	Precision in %	Recall in %	F1-Score in %
EfficientNet B0	80.133	81.89	80.67	80.34
EfficientNet B0 + Attention Block	86.67	87.612	87.65	86.98
EfficientNet B0 + Attention Block +LSTM	91.13	91.134	90.978	90.934

Table 4 — Performance of the proposed model for each fold

Fold number	Accuracy in %	Precision in %	Recall in %	F1-Score in %
Fold 1	91.25	91.70	91.89	91.67
Fold 2	90.80	90.89	90.55	90.55
Fold 3	91.18	91.78	91.03	91.03
Fold 4	90.56	90.06	90.12	90.12
Fold 5	91.86	91.24	91.3	91.3
Average	91.13	91.134	90.978	90.934

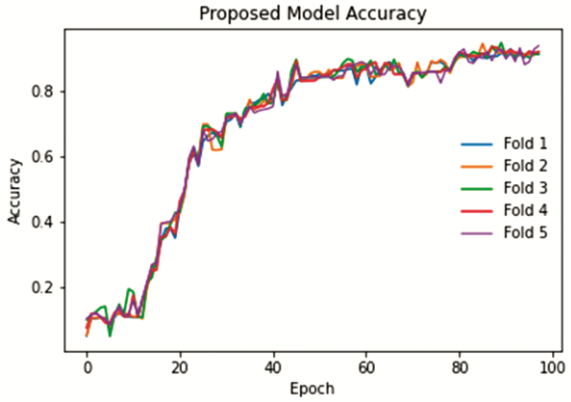

Fig. 7 — Observations of cross validation – Improvement in accuracy

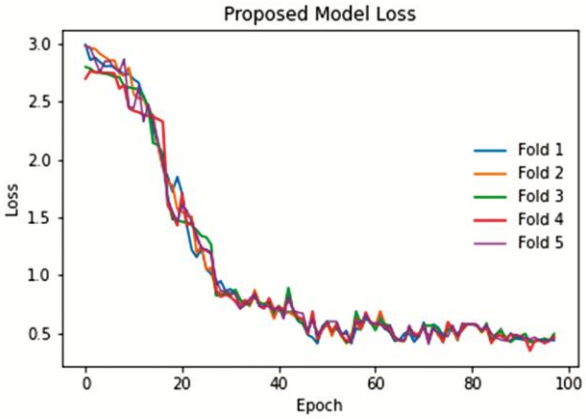

Fig. 8 — Observations of cross validation – Decay of loss

Precision, Recall, and F1-score in percent for every fold for the proposed model are presented in Table 4. The values obtained for the metrics used is shown in detail in Table 4.

The performance metrics such as Accuracy and Loss attained for every fold is compared and the related observations obtained are presented in Fig. 7 and Fig. 8 respectively.

Conclusions

This work analyses the performance of different deep learning architectures such as AlexNet, VGG16, InceptionV1, ResNet50, Auto-Encoder and EfficientNet B0 and eccentric methods such as Q-ADBN over long sequences of lip images of the MIRACL-VC1 dataset. The observations achieved from the models have indicated that EfficientNet B0 architecture yielded optimal performance with an accuracy of 80.133%. When, an Attention block is combined with EfficientNet B0 network architecture, the accuracy has gradually increased to 86.67%. The proposed model has LSTM layer coupled with EfficientNet B0 and Attention block and this amplified the accuracy of lip detection model and yielded an accuracy of 91.13%. The proposed model has integrated Attention block and LSTM layer with EfficientNet B0 neural network to improve the detection of lip movement. In future, the lip movement detection can be made extensive by using a diverse dataset with more words and phrases spoken by different speakers.

References

1 Qiao J, Wang G, Li W & Chen M, An adaptive deep Q-learning strategy for handwritten digit recognition, *Neural Netw*, **107** (2018) 61–71.

2 Rekik A, Ben-Hamadou A & Mahdi W, Human machine interaction via visual speech spotting, *Adv Concepts Intel Vision Syst*, (2015) 566–574.

3 Rekik A, Ben-Hamadou A & Mahdi W, Unified system for visual speech recognition and speaker identification, *Adv Concepts Intel Vision Syst*, (2015) 381–390.

4 Gergen S, Zeiler S, Abdelaziz A H, Nickel R & Kolossa D, Dynamic Stream Weighting for Turbo-Decoding-Based Audiovisual ASR, *Proc Interspeech*, (2016) 2135–2139.

5 Cooke M, Barker J, Cunningham S & Shao X, An audio-visual corpus for speech perception and automatic speech recognition, *J Acoust Soc Am*, **120(5)** (2006) 2421–2424.

6 Pei Y, Kim T & Zha H, Unsupervised random forest manifold alignment for lip reading, *Proc Int Comp Vis*, (2013) 129–136.

7 Ngiam J, Khosla A, Kim M, Nam J, Lee H & Ng A, Multimodal Deep Learning, *Proc Int Conf Mac Lear*, (2011) 689–696.

8 Assael Y M, Shillingford B, Whiteson S & de Freitas N, LipNet: End-to-end sentence-level lip reading, *ArXiv*, **(abs/1611.01599)**(2016).

9 Gutiérrez A & Alanah Robert Z, *Lip Reading Word Classification*, (2017).

10 Chung J S, Senior A, Vinyals O & Zisserman A, Lip Reading Sentences in the Wild, *Proc Int Comp Vis Patt*, (2017) 3444–3453.

11 Afouras T, Chung J S & Zisserman A, Deep Lip Reading: a comparison of models and an online application, *ArXiv*, **(abs/1806.06053)** (2018).

12 Wand M, Vu N T & Schmidhuber J, Investigations on end-to-end audiovisual fusion, *ArXiv*, **(abs/1804.11127)** (2018).

13 Shillingford B, Assael Y, Hoffman M, Paine T, Hughes C, Prabhu U, Liao H, Sak H, Rao K, Bennett L, Mulville M, Coppin B, Laurie B, Senior A & De Freitas N, Large-scale visual speech recognition, *ArXiv*, **(abs/1807.05162)** (2018).

14 Wang C, Multi-Grained Spatio-temporal Modeling for Lip-reading, *ArXiv*, **(abs/1908.11618)** (2019).
15 PetridisS, Shen J, Cetin D & Pantic M, Visual-only recognition of normal, whispered and silent speech, *Proc Int Acou Spec Recn* (2018) 6219–6223.
16 Rekik A, Ben-Hamadou A, Mahdi W, A new visual speech recognition approach for RGB-D cameras, in *Image Analysis and Recognition* edited by A Campilho, M Kamel, (ICIAR), *Lecture Notes in Computer Science*, **(8815)** (2014).
17 https://pypi.org/project/dlib/ (Accessed: 16 April 2021)
18 Tan M & Le Q V, EfficientNet: rethinking model scaling for convolutional neural networks, *ArXiv*, **(abs/1905.11946)** (2019).
19 Krizhevsky A, Sutskever I & Hinton G E, ImageNet classification with deep convolutional neural networks, *Commun ACM*, **60(6)** (2017) 84–90.
20 Simonyan K & Zisserman A, Very deep convolutional networks for large-scale image recognition (*ICLR*), *Proc Int Lear Rep*, 2015.
21 Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D & Vanhoucke V, Going deeper with convolutions, *Proc Int Comp Vis Patt*, (2015).
22 He K, Zhang X, Ren S & Sun J, Deep residual learning for image recognition, *Proc Int Comp Vis Patt*, (2016).
23 Hinton G E, Reducing the dimensionality of data with neural networks, *Science*, **313(5786)** (2006) 504–507.